Semantic-Consistent Dual-Image Vision-Language Alignment via Modality Routing for 3D PET/CT

Lin Lu[1], Hang Li[1], Zihan Liu[1], Chaoxiang Tang[1], and Hui Zhang[1]✉

[1] School of Biomedical Engineering, Tsinghua University, Beijing, China
✉ hzhang@tsinghua.edu.cn

Abstract. Positron Emission Tomography–Computed Tomography (PET/CT) provides complementary anatomical and metabolic information essential for cancer diagnosis. While vision-language pre-training has shown promise in medical imaging, existing approaches treat PET/CT as a unified modality, failing to capture the distinct characteristics of CT and PET imaging. We propose **PETCT-DiVLA**, a novel framework that introduces: (1) a token-level modality router that automatically identifies PET-specific, CT-specific, and shared text tokens; (2) a semantic-consistency guided SigLIP loss that leverages organ abnormality labels to define positive and negative pairs dynamically; and (3) dual-granularity alignment at both global and local levels. Our framework employs dual-branch 3D Vision Transformers and achieves mask-free inference. Experiments demonstrate state-of-the-art performance on image-text retrieval and classification tasks, including CT abnormality detection and PET uptake grading. Code is available at `https://github.com/Linn60/PETCT-DiVLA`.

Keywords: Vision-Language Pre-training · PET/CT Imaging · Medical Image Analysis

1 Introduction

Positron Emission Tomography–Computed Tomography (PET/CT) is widely used in modern oncology and routine clinical practice, providing complementary anatomical and metabolic information for cancer diagnosis, staging, and treatment monitoring [1]. While CT imaging captures high-resolution structural details, PET imaging reveals metabolic activity through radiotracer uptake. The fusion of these two modalities offers a comprehensive view of disease pathology. However, interpreting PET/CT studies requires specialized expertise, creating a practical bottleneck in clinical workflows.

Vision-Language Pre-training (VLP) has demonstrated remarkable potential in bridging medical imaging and natural language for representation learning [14,10]. Early approaches like ConVIRT [14] established foundations for medical contrastive learning in radiology, while subsequent works introduced multi-granularity alignment strategies [4,9]. Recently, specialized frameworks for PET/CT analysis have emerged [5,6].

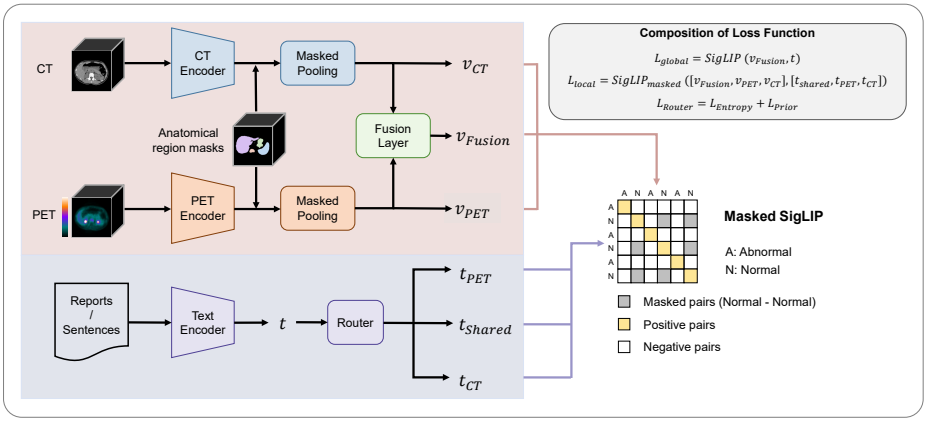

Fig. 1. The proposed PETCT-DiVLA framework consists of dual-branch 3D ViT encoders for PET and CT volumes, a text encoder with a token-level modality router, and dual-granularity alignment guided by semantic consistency.

However, vision-language alignment tailored to *dual-image* PET/CT inputs, especially in volumetric studies, remains relatively underexplored. For example, Jiao et al. [5] initialize the visual encoder from CT-pretrained weights and then train a vision-language model, which can bias representation learning toward CT-dominant structural cues during training and make it harder to capture PET-specific metabolic signals.

Despite these advances, existing approaches face fundamental limitations for PET/CT modeling. First, most methods treat PET/CT as a single concatenated input, failing to explicitly model the distinct information carried by each imaging modality. Second, conventional contrastive learning objectives suffer from the false negative problem, where semantically similar image-text pairs, such as shared normal findings, are incorrectly pushed apart. Third, alignment typically occurs at coarse granularities, missing fine-grained correspondences between anatomical regions and textual descriptions in reports.

To address these challenges within a unified PET/CT framework, we propose **PETCT-DiVLA** (Dual-Image Vision-Language Alignment for PET/CT). Our approach introduces three key innovations. First, we develop a token-level modality router that automatically identifies which text tokens describe PET-specific (metabolic), CT-specific (structural), or shared information. Second, we propose a semantic-consistency guided SigLIP loss that leverages organ-level abnormality labels to define more reliable positive and negative pairs dynamically. Third, we implement dual-granularity alignment at both global (whole-report–whole-body) and local (sentence–organ-level) scales within the same model.

Our contributions are: (1) Token-level modality routing for 3D PET/CT vision-language learning; (2) Semantic-consistency guided loss for mitigating false negative collisions; (3) Dual-granularity alignment with mask-free inference capability at test time.

2 Method

2.1 Overview

Our framework achieves semantic-consistent bimodal representation learning through a token-level modality router. Given a PET volume $\mathbf{P} \in \mathbb{R}^{H\times W\times D}$, a CT volume $\mathbf{C} \in \mathbb{R}^{H\times W\times D}$, and a radiology report $\mathcal{T} = \{t_1, \ldots, t_L\}$, our model learns joint representations in a shared embedding space.

2.2 Dual-Branch 3D Visual Encoder

Our visual encoder consists of two separate 3D Vision Transformers processing PET and CT volumes independently. Given input $\mathbf{X} \in \mathbb{R}^{H\times W\times D}$, we divide it into $N = 12 \times 12 \times 28 = 4032$ patches with size $(16, 16, 12)$:

$$\mathbf{F} = \mathrm{ViT}(\mathrm{PatchEmbed}(\mathbf{X})) \in \mathbb{R}^{N\times d} \tag{1}$$

Global visual representations are obtained via mean pooling for each modality:

$$\mathbf{v}_{\mathrm{pet}} = \mathrm{Norm}\left(\mathbf{W}_{\mathrm{pet}} \cdot \frac{1}{N}\sum_{i=1}^{N} \mathbf{F}_{\mathrm{pet}}[i]\right), \quad \mathbf{v}_{\mathrm{ct}} = \mathrm{Norm}\left(\mathbf{W}_{\mathrm{ct}} \cdot \frac{1}{N}\sum_{i=1}^{N} \mathbf{F}_{\mathrm{ct}}[i]\right) \tag{2}$$

The fused visual representation is:

$$\mathbf{v}_{\mathrm{fus}} = \mathrm{Norm}\left(\mathrm{fusion}([\mathbf{v}_{\mathrm{pet}}; \mathbf{v}_{\mathrm{ct}}])\right) \tag{3}$$

For local alignment, organ-specific features are extracted using masked pooling with segmentation masks M_o. Here, $M_o^{\downarrow}[i]$ denotes the patch-level organ mask value aligned with the ViT patch grid, obtained by downsampling the voxel-level mask to N patches:

$$\mathbf{v}_{\mathrm{pet}}^{o} = \mathrm{Norm}\left(\mathbf{W}_{\mathrm{pet}} \cdot \frac{\sum_{i=1}^{N} M_o^{\downarrow}[i] \cdot \mathbf{F}_{\mathrm{pet}}[i]}{\sum_{i=1}^{N} M_o^{\downarrow}[i] + \epsilon}\right) \tag{4}$$

$$\mathbf{v}_{\mathrm{ct}}^{o} = \mathrm{Norm}\left(\mathbf{W}_{\mathrm{ct}} \cdot \frac{\sum_{i=1}^{N} M_o^{\downarrow}[i] \cdot \mathbf{F}_{\mathrm{ct}}[i]}{\sum_{i=1}^{N} M_o^{\downarrow}[i] + \epsilon}\right) \tag{5}$$

2.3 Text Encoder with Token-Level Modality Router

We employ mmBERT-small [7] as our text encoder, producing contextualized token representations $\mathbf{H} \in \mathbb{R}^{L\times d}$. The global text representation is obtained via attention-masked mean pooling.

A central component of our text encoder is the token-level modality router, which computes a three-way distribution over $\{\mathrm{PET}, \mathrm{CT}, \mathrm{Shared}\}$:

$$\mathbf{w}_i = \text{Softmax}\left(\mathbf{W}_{\text{router}} \cdot \mathbf{h}_i\right) = [w_i^{\text{pet}}, w_i^{\text{ct}}, w_i^{\text{shr}}] \tag{6}$$

Soft pooling yields three text representations:

$$\mathbf{t}_{\text{pet}} = \text{Norm}\left(\frac{\sum_{i=1}^{L} w_i^{\text{pet}} \mathbf{h}_i a_i}{\sum_{i=1}^{L} w_i^{\text{pet}} a_i}\right) \tag{7}$$

with analogous expressions for $\mathbf{t}_{\text{ct}}$ and $\mathbf{t}_{\text{shr}}$.

2.4 Semantic-Consistency Guided Alignment

Global Alignment Loss. The global alignment enforces matching between fused visual and global text representations using SigLIP-style loss with learnable temperature τ and bias b:

$$\mathcal{L}_{\text{global}} = -\frac{1}{B}\sum_{i=1}^{B} \log \sigma(\text{sim}_{ii}) - \frac{1}{B(B-1)} \sum_{i \neq j} \log(1 - \sigma(\text{sim}_{ij})) \tag{8}$$

where $\text{sim}_{ij} = \tau \cdot \mathbf{t}_{\text{global}}^{(i)} \cdot \mathbf{v}_{\text{fus}}^{(j)} + b$.

Local Alignment Loss. The local alignment operates at the organ level with semantic-consistency constraints. Positive pairs are: (1) same patient and same organ, or (2) different patients, same organ, both normal. Negative pairs are excluded based on organ abnormality labels to avoid false negative collisions from template normal descriptions.

$$\mathcal{L}_{\text{local}} = \frac{1}{3|\mathcal{V}|} \sum_{k \in \{\text{pet,ct,shr}\}} \sum_{(i,j) \in \mathcal{V}} -\log \sigma\left(\ell_{ij} \cdot (\tau \mathbf{t}_k^{(i)} \mathbf{v}_k^{(j)\top} + b)\right) \tag{9}$$

where $\mathcal{V}$ denotes valid pairs and $\ell_{ij} \in \{+1, -1\}$ indicates positive/negative labels.

Router Regularization. Let $w_{l,k} \in [0,1]$ denote the router softmax weight of token l assigned to branch $k \in \{\text{pet}, \text{ct}, \text{shr}\}$, and $a_l \in \{0,1\}$ indicate whether token l is valid (from the attention mask). Let $\mathbf{w}_l = [w_{l,\text{pet}}, w_{l,\text{ct}}, w_{l,\text{shr}}]$. We regularize the router with a per-token entropy term and a global prior term:

$$\mathcal{L}_{\text{ent}} = -\frac{1}{L_{\text{valid}}} \sum_{l=1}^{L} a_l \sum_k w_{l,k} \log(w_{l,k} + \epsilon) \tag{10}$$

$$\mathcal{L}_{\text{prior}} = D_{\text{KL}}(\boldsymbol{\pi}_0 \| \boldsymbol{\pi}), \quad \boldsymbol{\pi} = \frac{1}{L_{\text{valid}}} \sum_{l=1}^{L} a_l \, \mathbf{w}_l, \quad \boldsymbol{\pi}_0 = [0.25, 0.25, 0.50] \tag{11}$$

where $L_{\text{valid}} = \sum_{l=1}^{L} a_l$ and ϵ is a small constant. Minimizing $\mathcal{L}_{\text{ent}}$ encourages more decisive (low-entropy) token-to-branch assignments, improving interpretability of PET/CT/shared text decomposition. Minimizing $\mathcal{L}_{\text{prior}}$ prevents routing collapse and enforces a desired global utilization ratio (load balancing) across the three branches.

2.5 Overall Training Objective

The total training loss combines all components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{global}} + \lambda_{\text{loc}}\mathcal{L}_{\text{local}} + \lambda_{\text{ent}}\mathcal{L}_{\text{ent}} + \lambda_{\text{pr}}\mathcal{L}_{\text{prior}} \tag{12}$$

with hyperparameters $\lambda_{\text{loc}} = 1$, $\lambda_{\text{ent}} = 0.1$, $\lambda_{\text{pr}} = 0.1$.

3 Experiments and Results

3.1 Datasets and Preprocessing

We evaluate on our private PET/CT dataset, which contains 8,155 paired PET/CT studies with matched Chinese radiological reports collected from Peking Union Medical College Hospital. The cohort consists of consecutive whole-body FDG PET/CT examinations collected over a 15-month period and covers multiple oncologic indications, reflecting a heterogeneous real-world PET/CT population from our institution. This study was approved by the Institutional Review Board of the participating institution. We use 7,355 cases for training and hold out 800 cases for testing. CT volumes are clipped to $[-1000, 1000]$ HU and PET to $[0, 20]$ SUV. All scans are resampled to a unified voxel spacing of $2 \times 2 \times 3$ mm^3 and then center-cropped/padded to $(192, 192, 336)$ for network input.

We apply TotalSegmentator [11] to the CT volumes to obtain segmentations of 28 anatomical regions: brain; orbits; nasal cavity/paranasal sinuses/nasopharynx; oral cavity/oropharynx; larynx/hypopharynx; thyroid and major salivary glands; cervical lymph nodes; lungs and pleura; mediastinum and hila; heart and pericardium; esophagus; axillae and chest wall soft tissue; breasts; liver; gallbladder and biliary tract; spleen; pancreas; stomach; bowel; kidneys; adrenal glands; ureters and bladder; reproductive organs; retroperitoneum; peritoneum/mesentery/omentum; pelvic and inguinal lymph nodes; skeletal system; other muscles and subcutaneous tissue. To derive structured supervision from free-text reports, similar to the PET/CT report structuring pipelines in [5,13], we use DeepSeek-V3.2 [2] to automatically parse each report and assign each sentence to an anatomical region, while labeling (i) whether CT findings are abnormal and (ii) the PET uptake grade. We aggregate these sentence-level predictions into region-level labels: CT abnormality indicators (normal or abnormal) and PET uptake grades (1–5): "1" (uptake defect/reduction), "2" (normal), "3" (physiological/baseline uptake), "4" (mild/suspicious abnormal uptake), "5" (obviously abnormal uptake).

3.2 Implementation Details

Our PETCT-DiVLA employs dual 3D ViT [3] encoders (dim=384, depth=12, heads=6) and mmBERT-small text encoder [7]. Training uses AdamW optimizer with learning rate 10^{-4}, batch size 16 per device, and cosine schedule with 4% warmup, trained on 8 NVIDIA A800 GPUs for 5,000 steps. Mixed-precision training (bfloat16) and gradient clipping (max norm = 1.0) are applied. To compute the local alignment loss, during training at each step we randomly sample 4 anatomical regions for each sample in the batch, while prioritizing abnormal regions to ensure informative contrastive signals.

3.3 Evaluation Metrics

For image-text retrieval, we report Recall@K (R@1, R@10, R@50) as percentages for both image-to-text (I2T) and text-to-image (T2I) directions. With 800 held-out cases, R@1 corresponds to retrieving the single correct match from 800 candidates.

All reported metrics are computed from the similarity between the fused visual embedding $\mathbf{v}_{\text{fus}}$ and the global text embedding $\mathbf{t}$ produced directly by the text encoder, without separately evaluating PET-only or CT-only image/text embeddings.

For classification, we report accuracy, F1-score, and area under the ROC curve (AUC). We perform region-wise evaluation over 28 anatomical regions, where each region is treated as an independent classification instance. Two classification tasks are evaluated: (1) CT abnormality detection (binary), and (2) PET uptake grading (5-class).

3.4 Image-Text Retrieval Results

Table 1 compares retrieval performance against baseline methods. Our method achieves superior retrieval performance, demonstrating the effectiveness of explicit modality handling and semantic-aware routing. Because retrieval is evaluated on the 800-case test set and sentences describing normal findings are often templated, absolute R@1 values are low; the consistent relative gains across R@K therefore better reflect improved alignment. For the CLIP and SigLIP baselines, we concatenate PET and CT into a two-channel image as the input to the visual encoder and continue training following their original training recipes.

3.5 Classification Results

Table 2 summarizes fine-tuned classification performance for both CT abnormality detection and 5-class PET uptake grading. For CT abnormality detection, PETCT-DiVLA achieves the best accuracy (0.562) and AUC (0.584), with AUC gains of 0.024, 0.039, and 0.023 over ViT, CLIP, and SigLIP, respectively. Although its F1-score is slightly lower than ViT, the higher AUC suggests better

Table 1. Image-Text Retrieval Performance Comparison. All R@K values are percentages.

Method	T2I R@1	T2I R@10	T2I R@50	I2T R@1	I2T R@10	I2T R@50
CLIP [8]	0.37	5.25	24.62	0.50	5.62	24.00
SigLIP [12]	0.37	6.00	26.12	0.50	6.25	26.25
PETCT-DiVLA	**1.25**	**8.38**	**28.87**	**1.37**	**7.87**	**29.62**

Table 2. Fine-Tuned Region-Wise Classification Performance

Method	CT Abnormality Detection			PET Uptake Grading		
	Acc.	F1	AUC	Acc.	F1	AUC
ViT [3]	0.557	**0.328**	0.560	0.759	0.731	0.573
CLIP [8]	0.534	0.321	0.545	0.735	0.709	0.541
SigLIP [12]	0.551	0.314	0.561	0.749	0.722	0.564
PETCT-DiVLA	**0.562**	0.319	**0.584**	**0.814**	**0.753**	**0.603**

overall separation between normal and abnormal regions. For PET uptake grading, PETCT-DiVLA consistently performs best, reaching 0.814 accuracy, 0.753 F1-score, and 0.603 AUC. The gains over the strongest baseline in each metric are 0.055, 0.022, and 0.030, indicating that explicit PET/CT routing better captures uptake-related semantics than simple two-channel concatenation.

3.6 Ablation Studies

Table 3 presents ablation results integrated with main metrics. The modality router contributes most significantly to performance, validating the importance of explicit PET/CT handling. Semantic consistency constraints provide substantial improvement by mitigating false negative collisions from template normal descriptions.

3.7 Qualitative Analysis

Fig. 2 provides t-SNE visualizations of the learned embedding spaces. Panel (A) plots the fused visual embeddings $\mathbf{v}_{\text{fus}}$ alongside global text embeddings $\mathbf{t}_{\text{global}}$. In this global space, all representations are projected without any modality-specific structure: PET-relevant and CT-relevant semantics are entangled together, offering no way to disentangle which information originates from metabolic imaging versus anatomical imaging. Panel (B) demonstrates the effect of our token-level modality router. After routing, four distinct groups emerge: $\mathbf{t}_{\text{PET}}$ and $\mathbf{v}_{\text{PET}}$ form a coherent cluster that is more separated from $\mathbf{t}_{\text{CT}}$ and $\mathbf{v}_{\text{CT}}$, while within each modality the visual and textual embeddings remain tightly aligned. This contrast highlights that our modality routing recovers fine-grained modality-specific structure on top of global matching, enabling the model to explicitly capture complementary PET metabolic and CT anatomical information that would otherwise be conflated in the undifferentiated global space of Panel (A).

Table 3. Ablation Study: Impact of Key Components

Configuration	T2I R@10	I2T R@10	PET AUC	CT AUC
L_{global}	6.88	6.25	0.531	0.544
$L_{global} + L_{local}$	7.25	7.50	0.591	0.575
Full Model (PETCT-DiVLA)	**8.38**	**7.87**	**0.603**	**0.584**

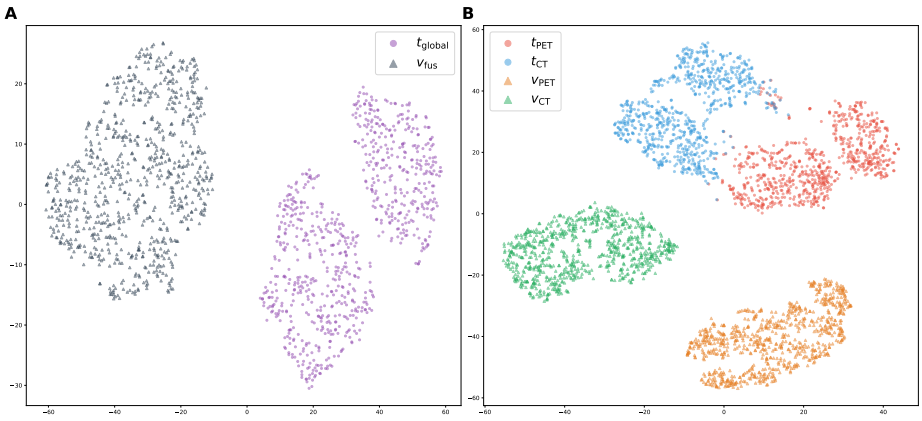

Fig. 2. t-SNE visualizations show the learned embedding spaces: (A) in the global embedding space, $\mathbf{v}_{\text{fus}}$ and $\mathbf{t}_{\text{global}}$ are entangled with no modality-specific structure; (B) after modality routing, PET and CT form distinct, well-separated subspaces with tight visual–text alignment within each modality, demonstrating the benefit of explicit modality decomposition.

Fig. 3 further visualizes token-level router behavior on a representative report. Words are colored by their router weights (red = PET, blue = CT, gray = shared): PET-related terms such as "radiotracer uptake" tend to appear red, whereas CT-related lesion terms such as "soft-tissue density opacities" tend to appear blue. Because these token-level weights are learned by the model rather than manually assigned, they are not accurate for every token, but a clear overall pattern still emerges. The colors are computed on the original Chinese report and mapped onto this English translation for readability, so each English word reflects the weight of its source Chinese token(s) rather than an English token.

4 Conclusion

We present PETCT-DiVLA for dual-image vision-language alignment in 3D PET/CT, combining token-level modality routing, semantic-consistency guided SigLIP training, and dual-granularity (global/local) alignment. PETCT-DiVLA achieves state-of-the-art results on image-text retrieval and classification, including CT abnormality detection and PET uptake grading, while preserving both fused and modality-specific representations. Future work will extend the framework to other multimodal imaging settings and report generation.

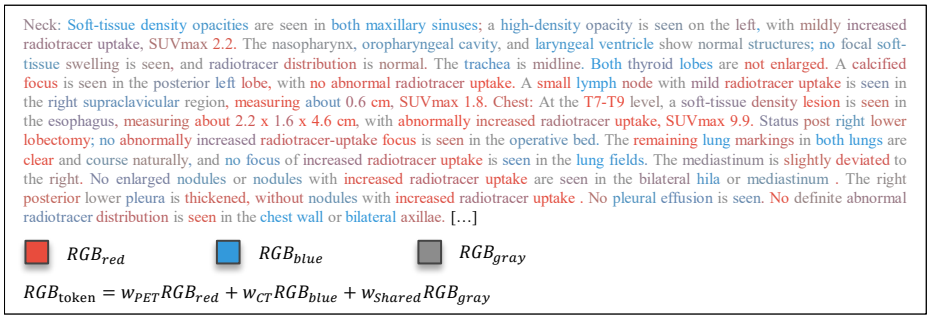

Fig. 3. Router weight visualization of a PET/CT report

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018)
2. DeepSeek-AI, Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., et al.: DeepSeek-V3.2: Pushing the frontier of open large language models (2025). `https://doi.org/10.48550/arXiv.2512.02556`
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
4. Huang, S.C., Shen, L., Lungren, M.P., Yeung, S.: GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3942–3951 (October 2021)
5. Jiao, W., Yan, K., Zhang, J., Jin, D., Xie, Z.: Vision-language models for automated 3d pet/ct report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5295–5304 (2026)
6. Maqbool, D., Lee, C., Huemann, Z., Church, S.D., Larson, M.E., Perlman, S.B., Romero, T.A., Warner, J.D., Lubner, M., Tie, X., et al.: Petar: Localized findings generation with mask-aware vision-language modeling for pet automated reporting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42637–42648 (2026)
7. Marone, M., Weller, O., Fleshman, W., Yang, E., Lawrie, D., Durme, B.V.: mmbert: A modern multilingual encoder with annealed language learning (2025)
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

9. Wang, F., Zhou, Y., WANG, S., Vardhanabhuti, V., Yu, L.: Multi-granularity cross-modal alignment for generalized medical visual representation learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 33536–33549. Curran Associates, Inc. (2022)
10. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: MedCLIP: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022)
11. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al.: Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
12. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11975–11986 (October 2023)
13. Zhang, Y., Zhang, W., Ling, Z., Feng, G., Peng, S., Chen, D., Liu, Y., Zhang, H., Wang, S., Li, L., et al.: Pet2rep: Towards vision-language model-drived automated radiology report generation for positron emission tomography. arXiv preprint arXiv:2508.04062 (2025)
14. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., Yeung, S. (eds.) Proceedings of the 7th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 182, pp. 2–25. PMLR (2022)